NUMERICAL METHODS IN TAXONOMY

R. C. Jancey*
Department of Biological Sciences, University of Sydney

[Read 24th November, 1965]

Synopsis

Some of the advantages of a numerical approach to taxonomy are indicated, also the compatibility of the technique with phylogenetically based taxonomy. Two main avenues for the application of computer techniques are described—the simplification of individual relationships and the detection of group structure. Finally, a means of combining the results of these two techniques is described.

The development of electronic computers has led to the introduction in recent years of numerical methods to the taxonomic process. That the advent of such methods has been the subject of some criticism cannot be denied, and it is the author's hope that this contribution will serve to dispel some misapprehensions, and to indicate some of the facilities offered by differing forms of numerical analysis.

Perhaps the foremost objection of many taxonomists to the introduction of numerical methods is their doubt that any automatic process could replace the extremely complex and flexible mental comparison of individuals and attributes which forms the vital part of the taxonomic process. The second objection is that the use of numerical methods precludes any phylogenetic basis for the final classification, and is hence an essentially retrogressive step. With regard to the first objection, assurance may be given that computers are indeed capable of reproducing the results of mental classifications made by taxonomists, so long as they are provided with the same or comparable information. A number of methodological investigations of numerical techniques have been carried out in which data supplied by monographers have been subjected to numerical analysis, the resulting output being fully in accord with the taxonomic decisions arrived at independently by the monographer (Rogers and Fleming, 1964). In an investigation of the genus *Phyllota* (Leguminosae) the author demonstrated by numerical methods a group structure almost identical with one advocated by Bentham more than a century ago, although in this case the characters used were almost certainly quite different (Jancey, 1965).

The second objection to the use of numerical methods, that they preclude a phylogenetic classification, is quite unfounded though widely held. Such a situation may well be the result of a misunderstanding since, although a number of numerical taxonomists hold strong views on the place of phylogenetic considerations in classification, this in no way makes numerical techniques and a phylogenetic classification necessarily incompatible. Assuming that the term phylogenetic classification implies the interpretation and modification of the groupings of present day phenotypically similar organisms in the light of known or inferred evolutionary trends, then a number of observations may be made concerning the methods by which such a classification can be achieved. In the mental taxonomic process it is possible to keep evidence of evolutionary trends in mind, and to modify taxonomic relationships even as they are being

* Author's present address: The Department of Quantitative Taxonomy, New York Botanical Garden, Bronx 58, New York, U.S.A.

constructed, or, alternatively, to construct first an essentially phenotypic classification and then modify this in the light of such other evidence as may be available. Both these possibilities are available with numerical techniques; in the case of concurrent consideration of evolutionary evidence, such data would have to be converted into a subjective numerical form, its relative influence on the final result being entirely in the hands of the taxonomist during the coding process. If it should be asked how one estimates the importance of a hypothetical trend, relative to a given piece of phenotypic data, it may be pointed out that such a subjective estimate must be made, at least subconsciously, in the mental process, and that an attempt to arrive at a visible estimate of such degrees of relative importance could in itself be illuminating. Such a mixing of the factual with the hypothetical is the basis for the objection of numerical taxonomists to the inclusion of evolutionary data in the analyses themselves. The second approach indicated above is perhaps the more satisfactory; the consideration of phylogenetic data after completion of a purely phenotypic grouping would result in the final phylogenetic classification being the same, but it would then be possible to see precisely what changes in a purely phenotypic grouping had been made by the taxonomist in order to achieve a more phylogenetic relationship, thus opening the way for a more informed discussion of the significance of such changes.

Apart from their acceptability as techniques for performing the sorting and group-forming processes of taxonomy, numerical analyses offer a number of additional benefits. Information concerning the homogeneity and relative similarity of the groups is available from most analytical methods, thus enabling the purely taxonomic decisions regarding the status of the groups to be based on rather more precise evidence than usual. The analytical techniques involved are mathematically defined and reproducible, thus the repetition of an analysis with new or differently defined characters is capable of providing additional evidence concerning the validity of the original choice of characters, or classification arrived at, since the computational procedure itself remains constant.

The basis of numerical methods

The taxonomic process is essentially the translation of observations made on individuals into statements of similarity and hence of group structure. The use of mathematical techniques in taxonomy has been largely confined in the past to their secondary applications of describing and substantiating taxa which have been established by subjective processes. Techniques of this type, e.g. Analysis of Variance, Discriminant Functions, still require the prior establishment of groups by some means or other before they can be applied. It is only with the advent of electronic computers that it has become practicable to carry objective translations of information concerning individuals into statements of group structure. While such analyses almost all start by computing some measure of similarity between all possible pairs of individuals, they differ greatly in the way in which this information is used to detect group structure.

One of the first results of the use of numerical methods of data analysis is an increased realization of the multidimensional nature of taxonomic relationships. A single dimension is sufficient to describe the relationships of two points. If a third point is added, a statement of its relationship to the first point will necessarily fix its position relative to the second point, a position which may well not represent its true relationship. This is a difficulty which may be overcome by the addition of a second dimension. Clearly, the addition of a fourth point may require the addition of a third dimension, so that in general terms it may be said that n–1 dimensions will be needed to describe fully all possible relationships of n points. This statement will be obviously as true for taxa or individuals as for points, though it must be emphasized that this is the maximum number of dimensions which may be needed, particular cases

may well require fewer, the simplest situation being a straightforward clinal variation with n points arranged in a straight line.

It should be pointed out before proceeding further that not all analytical methods employ the multidimensional Euclidean space foreshadowed above. Indeed, the advantages of an entirely non-metric space for the detection of group structure are considerable (Rogers and Fleming, 1964). It is felt, however, that the concept of similarity between individuals or groups is intuitively considered in terms of real spatial relationships, and that for this reason the concept is worthy of retention and consideration in greater detail, even though it involves excursions beyond three dimensions.

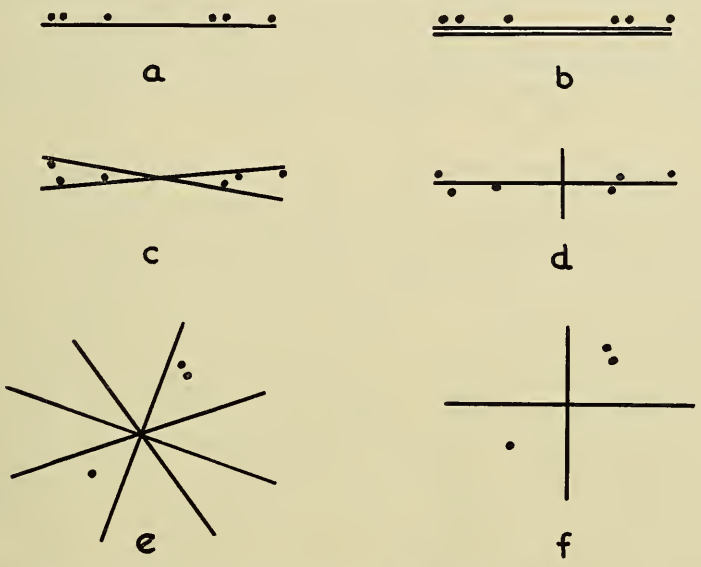

Fig.1. The representation of correlated characters by oblique axes.

a, Six individuals arranged according to their mutual phenotypic similarities, as revealed by a single character. *b*, As in *a*, but with two perfectly correlated characters. *c*, The same six individuals, now showing their phenotypic similarities as revealed by two highly, but not perfectly correlated characters (the cosine of the angle enclosing the points is equal to the correlation between the characters). *d*, As in *c*, but with two orthogonal axes now replacing the two oblique axes of the correlated characters. The positions of the points remain unchanged relative to each other. *e*, Three individuals described in terms of four correlated characters (the four axes are not necessarily confined to two dimensions). *f*, As in *e*, but re-expressed without loss of information in terms of two orthogonal axes (the maximum number needed to express the relationships of three individuals).

Information as collected by the taxonomist is expressed in terms of a number of reference variables, i.e. the characters observed and recorded for each specimen, the variables being more or less correlated. Thus the phenotypic relationships of a collection of individual plants for which x characters have been recorded may be thought of as being described in terms of x oblique axes (oblique because of the character correlations), the axes being located in a space of at most $n-1$ dimensions where n equals the number of individuals included in the analysis. Thus the relationships of the individual specimens could equally well be represented by $n-1$ orthogonal axes as by the x oblique ones. If x is less than $n-1$ then x represents the maximum number of dimensions required to represent the information available, the extent to which this number of dimensions can be reduced depending on the extent to which the characters are correlated (see Fig. 1).

While $n-1$ dimensions represent the maximum number of dimensions needed for complete description of the population, in practice far fewer

dimensions are needed to contain the information available. Indeed, a further reduction in the number of dimensions may be achieved with a level of distortion which would be quite acceptable in the interests of simplicity of description. Thus one of the main objects of a taxonomic computer programme is to take a population whose interrelationships are described in terms of a large number of correlated characters, and to re-express the interrelationships in terms of a relatively small number of dimensions, while allowing the taxonomist to nominate the level of distortion, if any, which is acceptable. At the same time, and unlike classical taxonomic methods, the processes to which the data have been subjected are completely definable. There are then four further steps in the taxonomic process, all of which will have been simplified by the re-expression of the characters. Firstly, an examination of the spatial relationships of individuals for evidence of group structure; secondly, the assignment of individuals to groups; thirdly, an evaluation of the relationships between the groups; and finally, the setting up of characters, or linear compounds of characters (cf. discriminant functions) to discriminate between the groups.

The technique of factor analysis is particularly well adapted to performing the first part of this process. It is a technique first used by Spearman (1904 et seq.) to describe the results of a large number of different tests of human ability in terms of a relatively small number of special aptitudes, e.g. manual, visual, numerical, etc., each special aptitude being described by a linear compound of the original tests. This is clearly the same as the first part of the taxonomic process, and by a slight extension, can be used as such. The analysis is based on the formation of a correlation or similar matrix from the original characters, from which is extracted a series of vectors or factors compounded from the characters. Since these factors when extracted from the matrix are made up of varying contributions from the original characters, it is possible to re-state the population relationships in terms of factor scores rather than characters (for a full account of factor analysis, see Harman, 1960). The relative information content of the factors depends on the method used for extracting them from the matrix, and for taxonomic purposes particular requirements for information distribution apply. The prime purpose of factor analysis of taxonomic data, as has been stated, is to reduce as far as possible the number of dimensions used in taxonomic description, so that as small as possible a number of meaningful factors is desirable. The Principal Axes method of factor analysis is such that the residual variance of the matrix is minimized with the extraction of each factor, thus the first factor extracted will contain the most information, and although in the Principal Axes method the number of factors extracted is equal to the order of the original correlation matrix, the information content of succeeding factors falls off rapidly and becomes non-significant. In graphic terms, the analysis examines a population described in terms of a number of oblique axes set in a multidimensional space, and computes the one axis best able to describe the spatial relationships of the population, the axis being composed of a linear compound of the original characters used. The analysis then investigates the position of the axis best able to represent the spatial relationships undescribed by the first axis. By definition, these axes and the succeeding ones must be orthogonal to each other. Knowing the contributions of the original characters to each factor, it is possible to re-express the data in terms of factors. By expressing the relationships of the individuals in terms of the first three factors only, a loss of information is incurred, but because of the rapid fall off in information content of the factors this is not usually serious, but has the advantage that limitation to three factors enables the spatial relationships of the individuals to be expressed graphically using isometric graph paper.

While factor analysis does not, in itself, delimit groups, it does present data in a far more comprehensible form as a basis for the establishment of such

groupings by other means. Methods are available which do make an objective demonstration of group structure, notably those of Goodall (1953), Michener and Sokal (1957), Sneath (1957), Williams and Lambert (1959), and Rogers and Fleming (1964). A fuller account of these methods will be found in Sokal and Sneath (1963), but for the purposes of this discussion it is sufficient to say that while considerable differences exist between the respective techniques, they all depend essentially on the calculation of some measure of association between all possible pairs of individuals, based on the characters measured. Such a measure of association can take many forms, being usually based on the ratio of character matches to mismatches between pairs of individuals, since such a measure is particularly well suited to data in a presence or absence, or limited class form. Having computed some measure of association, groups may be established by a synthetic process, the agglomeration of individuals possessing mutually high levels of association, discontinuities in the agglomerative process indicating group structure. The actual delimitation of groups may be performed automatically in response to some parameter involving the discontinuities—essentially a relationship between variation within the group and that of the whole population.

The techniques of Goodall (1953) and of Williams and Lambert (1959) are rather different in that they are analytic processes designed primarily for ecological use, whereby the population is subjected to successive divisions into the most homogeneous sub-groups. Such methods are particularly adapted to two-state character data, and provide a monothetic classification with hierarchical ordering of groups.

The methods of detecting group structure which have been described above do not by themselves give any obvious indications of the inter-relationships of the groups demonstrated. Levels of similarity at which groups form from individuals submitted to analysis are usually shown in the form of a dendrogram. Such diagrams have the advantage of illustrating clearly the discontinuities between groups. They cannot, however, represent in graphic form the similarity relationships between all pairs of individuals. Such a representation is not possible in two dimensions for the reasons described previously. A pictorial representation approximating to group inter-relationships may be obtained by the combination of factor analysis with one of the techniques of group detection described. The centres of gravity of the groups can be calculated in terms of three-dimensional space from the factor scores of individuals on the first three axes of a factor analysis. Knowing the individual co-ordinates of members of a group, a value for the standard deviation from the mean can be calculated for the group on each axis. It is thus possible to construct a perspective diagram of the group relationships on isometric graph paper, in which the groups are represented as ellipses drawn at one standard deviation from the mean of the group on each factor axis, the ellipses serving to indicate both the spatial relationships and the amount of variation found within and between the groups. It might be argued that more real information could be obtained from a perspective diagram showing the positions of all the individuals on which the analysis was based. Such a diagram is impracticable, since the illusion of three dimensions is lost when a large number of points need to be shown, and in addition no advantage would have been gained from the objective discrimination of groups made previously. An example of a perspective diagram of the former type is shown in Jancey (1966).

Conclusions

Numerical methods of data analysis are considered by the author to represent a valuable new technique available to the taxonomist. It is unfortunate that some taxonomists have looked upon the technique as an isolated field of endeavour, together with chemotaxonomy and cytotaxonomy,

and bearing no close relation to taxonomy as practised in the herbarium. While for purely practical reasons classification must continue to be based largely on morphological data, it would seem unreasonable to ignore any additional information concerning the living organisms which might be available. Similarly, a technique which enables the maximum amount of information to be extracted from a mass of raw data by a defined process would seem to be worthy of consideration by all taxonomists. The nature of the results yielded by numerical methods should be emphasized, since they are a frequent source of misunderstanding. The computations do not produce classical taxa, but group the individuals for which data was provided. Precise information is provided concerning the membership, distinctness, and diagnostic characters of the groups produced, but the status of any group in terms of orthodox taxonomic nomenclature, and its relationship to other taxa, are entirely in the hands of the taxonomist, the only difference being that he is provided with rather more information than usual on which to base his decision.

References

Goodall, D. W., 1953.—Objective methods for the classification of vegetation. (1). *Austr. Journ. Bot.*, 1 : 39–63.

Harman, H., 1960.—" Modern Factor Analysis." (University of Chicago Press, Chicago.)

Jancey, R. C., 1966.—An investigation of the genus *Phyllota* (DC.) Benth. Proc. Linn. Soc. N.S.W., 90 : 341–375.

————, 1965.—The application of numerical methods of data analysis to the genus *Phyllota* (DC.) Benth. *Aust. Journ. Bot.* (in press).

Michener, and Sokal, R. R., 1957.—A quantitative approach to a problem in classification. *Evolution*, 2 : 130–162.

Rogers, D. J., and Fleming, H., 1964.—A computor program for classifying plants (II). *Bio. Science*, 14 : 9 : 15–28.

Sneath, P. H. A., 1957.—The application of computors to taxonomy. *Journ. Gen. Microbiol.*, 17 : 201–226.

Sokal, R. R., and Sneath, P. H. A., 1963.—" Principles of Numerical Taxonomy." (W. H. Freeman and Co., San Francisco and London.)

Spearman, C., 1904.—General intelligence objectively determined and measured. *Amer. J. Psychol.*, 15 : 201–293.

Williams, W. T., and Lambert, J. M., 1959.—Multivariate methods in plant ecology (I). Association analysis in plant communities. *J. Ecol.*, 47 : 83–101.